

# Minimum Probe Length for Unique Identification of All Open Reading Frames in a Microbial Genome

*B. A. Sokhansanj, J. Ng, J. P. Fitch*

This article was submitted to 8<sup>th</sup> International Conference on Intelligent Systems for Molecular Biology, La Jolla, CA, August 20 – 23, 2000

U.S. Department of Energy

Lawrence  
Livermore  
National  
Laboratory

**March 5, 2000**

## DISCLAIMER

This document was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor the University of California nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or the University of California, and shall not be used for advertising or product endorsement purposes.

This is a preprint of a paper intended for publication in a journal or proceedings. Since changes may be made before publication, this preprint is made available with the understanding that it will not be cited or reproduced without the permission of the author.

This report has been reproduced directly from the best available copy.

Available electronically at <http://www.doc.gov/bridge>

Available for a processing fee to U.S. Department of Energy  
And its contractors in paper from  
U.S. Department of Energy  
Office of Scientific and Technical Information  
P.O. Box 62  
Oak Ridge, TN 37831-0062  
Telephone: (865) 576-8401  
Facsimile: (865) 576-5728  
E-mail: [reports@adonis.osti.gov](mailto:reports@adonis.osti.gov)

Available for the sale to the public from  
U.S. Department of Commerce  
National Technical Information Service  
5285 Port Royal Road  
Springfield, VA 22161  
Telephone: (800) 553-6847  
Facsimile: (703) 605-6900  
E-mail: [orders@ntis.fedworld.gov](mailto:orders@ntis.fedworld.gov)  
Online ordering: <http://www.ntis.gov/ordering.htm>

OR

Lawrence Livermore National Laboratory  
Technical Information Department's Digital Library  
<http://www.llnl.gov/tid/Library.html>

# Minimum Probe Length for Unique Identification of All Open Reading Frames in a Microbial Genome

Bahrad A. Sokhansanj, Jefferson Ng, J. Patrick Fitch

Biology & Biotechnology Research Program  
Lawrence Livermore National Laboratory  
L-452, 7000 East Avenue  
Livermore, CA 94550  
tele. (925) 422-8643, fax (925) 423-3068  
[sokhansanj@llnl.gov](mailto:sokhansanj@llnl.gov), [ng21@llnl.gov](mailto:ng21@llnl.gov), [jpfitch@llnl.gov](mailto:jpfitch@llnl.gov)

**Keywords:** oligonucleotide probes, complete genome, DNA microarrays, microbial genomics

## Abstract

In this paper, we determine the minimum hybridization probe length to uniquely identify at least 95% of the open reading frame (ORF) in an organism. We analyze the whole genome sequences of 17 species, 11 bacteria, 4 archaea, and 2 eukaryotes. We also present a mathematical model for minimum probe length based on assuming that all ORFs are random, of constant length, and contain an equal distribution of bases. The model accurately predicts the minimum probe length for all species, but it incorrectly predicts that all ORFs may be uniquely identified. However, a probe length of just 9 bases is adequate to identify over 95% of the ORFs for all 15 prokaryotic species we studied. Using a minimum probe length, while accepting that some ORFs may not be identified and that data will be lost due to hybridization error, may result in significant savings in microarray and oligonucleotide probe design.

## Introduction

Arrays of cDNA spotted on glass or nylon, or synthesized on silicon or glass chips have been developed in the last decade to simultaneously measure the expression of thousands of different genes by hybridization. These genes may be subsets of eukaryotic genome responsible for specific cellular functions, such as DNA repair, or they may even represent the entire genome of a microbial organism ( $\sim 10^3$  genes). In the simplest case the entire DNA sequence of the gene is used as the array. However, some of these cDNAs would have to be thousands of bases in length, and there are practical problems with implementation. A more efficient approach is to use a shorter cDNA probe that represents a subsequence unique to the gene. For example, the sequences "ACTCG" and "ACTCT" may be uniquely identified by the probes "TCG" and "TCT".

Unfortunately, DNA and RNA do not hybridize perfectly. The stacking interactions between bases in a sequence are so stabilizing that a few base pair mismatches over hundreds or thousands of bases are not thermodynamically unfavorable. Indeed, cells contain

elaborate DNA repair mechanisms to repair errors arising from these mismatches. To avoid hybridization errors, chip and array designers typically design specialized probe sets that include as much unique sequence for each gene as possible. They also include controls that indicate to the experimenter that hybridization errors may have occurred. As a result of this, a 25-mer probe is to date the shortest practical length, and chip and array manufacturers use sophisticated proprietary algorithms to design probe sets (Lipshutz et al. 1999).

But, if we sacrifice some accuracy and find a way to avoid hybridization errors, a less costly and more convenient approach may become feasible. The number of permutations of a sequence of  $N$  bases is  $4^N$ . For  $N = 10$ , there are more than  $10^6$  possible 10-mers. At first glance, it appears likely that each of  $10^3$  genes contain at least one unique 10-mer subsequence. This 10-mer could then be used to uniquely identify the gene's presence in a hybridization experiment. This has been the extent of the analysis in previous literature about using very short probes to identify genes (Velculescu et al. 1995).

The problem is more complicated, however. If genes have very similar sequences they will be fewer distinct subsequences. Fortunately, structural genes deviate by virtue of coding for different amino acid sequences. Unfortunately, this is not true of all genes, particularly those that regulate expression. Hopefully, there may still be a probe length  $N$  for which 90% or more of the genes can be identified.

In this paper, we determine the minimum probe length required to identify at least 90% of predicted open reading frames (ORFs) in the whole genome sequences of nine microbial species and two eukaryotic species (the yeast *S. cerevisiae* and the nematode *C. elegans*). We accomplish this by searching through publicly available sequence data to count the number of uniquely identified ORFs for different subsequence lengths. We also show that a simple rule-of-thumb model that assumes genes are random IID

sequences correctly predicts the same minimum probe length, but also implies unique identification of all genes, which is generally not true for a real genome.

## Scanning Whole Genome Sequences

We downloaded publicly available whole genome sequences of 17 species. In all cases, we used the open reading frames (ORFs) found by the local site from which the data was downloaded. Table 1 lists the species used in this study, along with kingdom and sequencing group.

Species	ORFs	Reference	Site
<i>Yersinia pestis</i>	B 4060	Sanger Centre, unpubl.	1
<i>Escherichia coli</i>	B 4405	Blattner et al. 1997	2
<i>Bacillus subtilis</i>	B 4094	Kunst et al. 1997	3
<i>Campylobacter jejuni</i>	B 1731	Sanger Centre, unpubl.	1
<i>Thermatoga maritima</i>	B 1877	Nelson et al. 1999	4
<i>Deinococcus radiodurans</i>	B 3187	White et al. 1999	4
<i>Borellia burgdoferi</i>	B 1738	Fraser et al. 1997	4
<i>Haemophilus influenzae</i>	B 1738	Fleischmann et al. 1995	4
<i>Helicobacter pylori</i>	B 1590	Tomb et al. 1997	4
<i>Trepanoma pallidum</i>	B 1039	Fraser et al. 1998	4
<i>Mycoplasma genitalium</i>	B 483	Fraser et al. 1995	4
<i>Pyrococcus horokoshii</i>	A 2061	Kawarabayasi et al. 1998	5
<i>Pyrococcus abyssi</i>	A 1816	CNS Genoscope, unpubl.	6
<i>Methanococcus janaschii</i>	A 1783	Bult et al. 1996	4
<i>Archaeoglobus fulgidus</i>	A 2437	Klenk et al. 1997	4
<i>Saccharomyces cerevisiae</i>	E 6085	Goffeau et al. 1996	7
<i>Caenorhabditis elegans</i>	E 15991	C. Eleg. Seq. Cons. 1998	7

B.: Eubacteria

A.: Archae

E.: Eukarya

1. <http://www.sanger.ac.uk>

2. <http://www.genetics.wisc.edu>

3. <http://bioweb.pasteur.fr>

4. <http://www.tigr.org>

5. <http://www.bio.nite.go.jp>

6. <http://www.genoscope.cns.fr>

7. <http://www.ncbi.nlm.nih.gov>

**Table 1 Genomes downloaded for analysis**

We downloaded sequence data already divided into ORFs. Results will vary based on different ORFs predicted by different finding algorithms. However, finding prokaryotic ORFs is relatively straightforward because a single gene is not divided into multiple, discontinuous ORFs. The actual ORFs should not differ sufficiently from the predictions enough to change our conclusions.

We determined the number of uniquely identifiable ORFs for a particular probe length  $N$  as follows. We considered each ORF as a separate string of characters (the bases 'A', 'C', 'T', 'G') and assigned the ORF a unique index. Starting with the first ORF (index 1), we defined a substring of length  $N$  from the first character to the  $(N-1)$ th character of the string. The substring was then inserted in a lookup table corresponding to the ORF index 1. Then, we defined a substring from the second character to the  $N$ th character, and again place it in the lookup table assigned to the ORF index 1. This continued until the last complete substring of length  $N$  was extracted from the first ORF string. Then we proceeded to the second ORF, index 2, and considered all of its constituent substrings. If a substring was not found in the lookup table, it was inserted in the table and assigned to ORF index 2. If it was found in the lookup table, it was assigned a negative index, indicating that it was not uniquely found on an ORF. We systematically searched all ORF strings in this fashion. Finally, we searched the lookup table, and counted the number of different positive ORF indices left assigned to a substring.

The algorithm was repeated for probe substring lengths  $N = 7$  to 14. It was implemented using Perl 5.0 on a Sun Ultra 5, and took approximately one hour to completely search through a microbial genome containing on the order of  $10^3$  ORFs.

## Using a Simple Model to Predict Minimum Probe Length

We construct a simple mathematical expression for the expected number of ORFs that contain at least one unique subsequence probe of given length  $N$ . We will assume that ORFs are random sequences of *constant* length  $L$  consisting of bases A, C, T, and G occurring in *equal* concentrations. Previous work indicates that a random model may succeed, because DNA sequences have much higher entropy and are consequently more disordered and random than other kinds of information, like language and music [16]. We also assume that every subsequence (including overlaps) of an ORF sequence is *distinct*; that is, every ORF has the maximum  $L-N+1$  different subsequences.

Under these assumptions, given  $G$  random ORF sequences, the probability that a particular ORF sequence contains at least one subsequence (s.s) of length  $N$  that

does not occur within any other arbitrary ORF sequence is:

$$\begin{aligned}
P(\geq 1 \text{ unique s.s.}) &= 1 - P(0 \text{ unique s.s.}) \\
P(0 \text{ u.s.s.}) &= P(1st \text{ s.s. not unique}) \times P(2nd \text{ s.s. not unique}) \times \dots \times P((L - N + 1)th \text{ s.s. not unique}) \\
&= \prod_{k=1}^{L-N+1} P(kth \text{ s.s. not unique}) \\
&= P(\text{random s.s. not unique})^{L-N+1} \\
\text{but,} \\
P(s.s. \text{ not unique}) &= 1 - P(s.s. \text{ not found in any other sequence}) \\
&= 1 - P(s.s. \text{ not found in random sequence})^{G-1}
\end{aligned}$$

Thus, we define the probability  $P_U$  that an ORF is uniquely identified by at least one probe of length  $N$ :

$$P_U(N, L, G) = 1 - \left(1 - P_0(N, L)\right)^{G-1} \quad (1)$$

where  $G$  is the total number of ORFs in the genome,  $N$  is the subsequence probe length, and  $P_0(N, L)$  is the probability that an arbitrary random sequence of length  $L$  does not contain a particular subsequence of length  $N$ .

For given  $N$ , we can calculate  $P_0$  using the recursion equation (Bloom 1996):

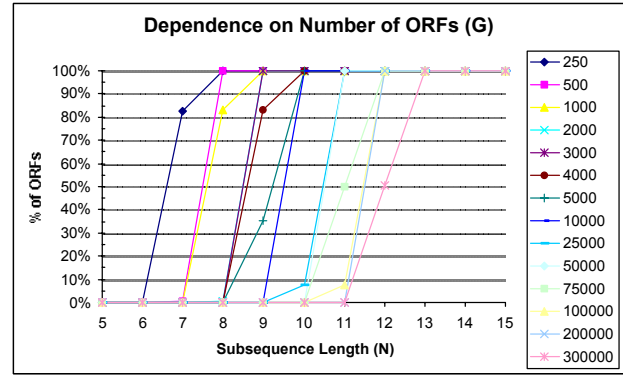
$$\begin{aligned}
P_0(L) &= P_0(L-1) - p^N (1-p) P_0(L-N-1); N > L \\
P_0(0) &= P_0(1) = \dots = P_0(N-1) = 1; N < L \quad (2) \\
P_0(N) &= 1 - p^N; N = L
\end{aligned}$$

The recursion (2) will give a binomial distribution. Since  $P_0$  dominates the result of (1), we expect that when the proportion of ORFs with a unique probe  $P_U$  is plotted for different values of  $N$ , it will look like a characteristic sigmoid. There will be a critical probe length  $N_C$  at which there is a sudden increase in the number of uniquely identifiable ORFs.

Fig. 1 shows this plot for different numbers of ORFs within a genome, assuming a constant ORF sequence length of 500 and equal base distribution.

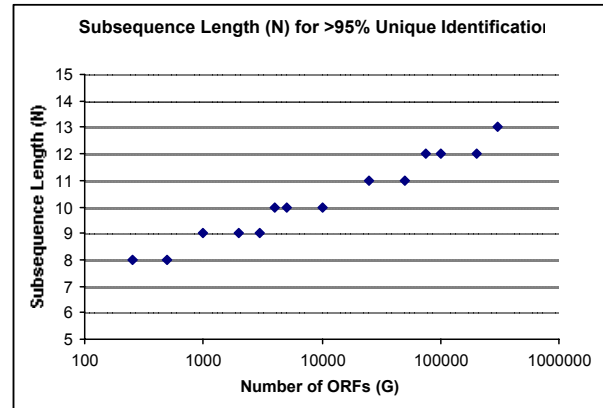
Based on our simple model, for microbial genomes where  $G \sim 10^3$  ORFs, the minimum probe length for most ORFs to be identified is 9 or 10 bases. We tested the model by comparing the result of (1) with the results of the exhaustive search for each of the species, using the species genome size as the input  $G$ . Throughout, we assumed a constant ORF length  $L$  of 500 and an equal base distribution ( $p = 0.25$ ). These are arbitrary but

realistic values intended to avoid reparameterizing the expression for every different species.



**Figure 1** Proportion of uniquely identifiable ORFs for different genome sizes, ORF length of 500 bases

Fig. 2 shows a plot of the critical subsequence probe length  $N$  at which 95% or more of the ORFs are uniquely identifiable, with the number of ORFs on a log axis. Again, this is for a constant ORF Length of 500 and equal base distribution.



**Figure 2** Minimum probe length as genome size increases

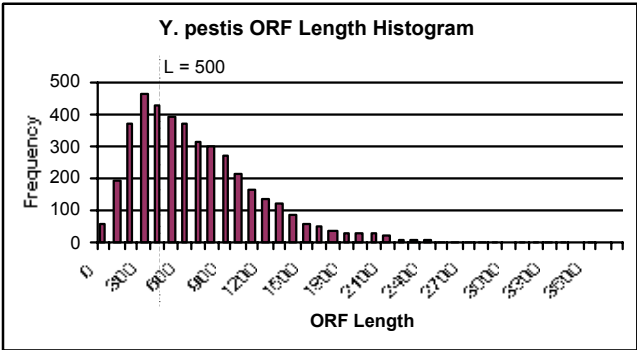
**Assumption of Equal Base Distribution.** Table 2 shows the variability in base concentrations for three particular species.

	A	T	G	C	A-T	G-C
<i>M. genitalium</i>	36%	32%	17%	15%	68%	32%
<i>E. coli</i>	24%	24%	27%	24%	48%	52%
<i>Y. pestis</i>	25%	26%	26%	23%	51%	49%

**Table 2** Distribution of bases in ORF data

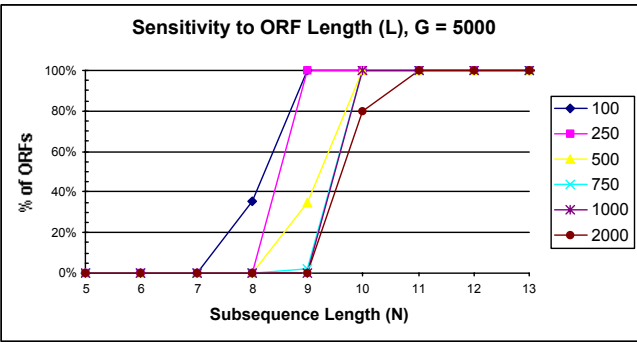
For many prokaryotic genomes, assuming equal base distribution is appropriate. This will not typically be the case for eukaryotic genomes for which G/C dominates in ORFs.

**Assumption of Equal ORF Length.** Fig. 3 shows a histogram demonstrating the large variation in lengths for *Y. pestis*. While  $L = 500$  may be lower than the mean length, it is a convenient constant value that is well within the peak of the distribution.



**Figure 3** ORF length histogram of *Y. pestis*

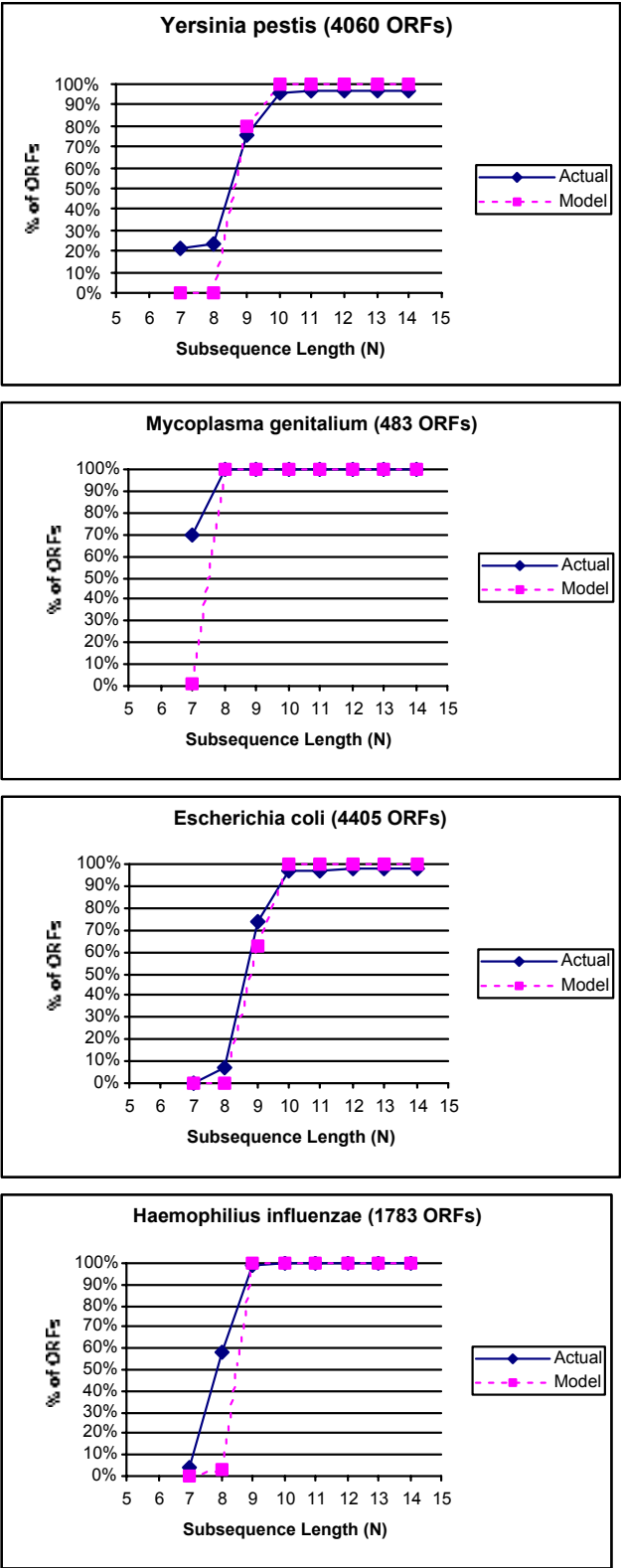
Also, as Fig. 4 shows, the choice of  $L$  does not significantly affect the model’s prediction of uniquely identifiable ORFs. For a genome size of 5000, typical for prokaryotic species, changing the ORF length from 500 through 1000 bases does not change the subsequence length  $N$  for 100% identification.

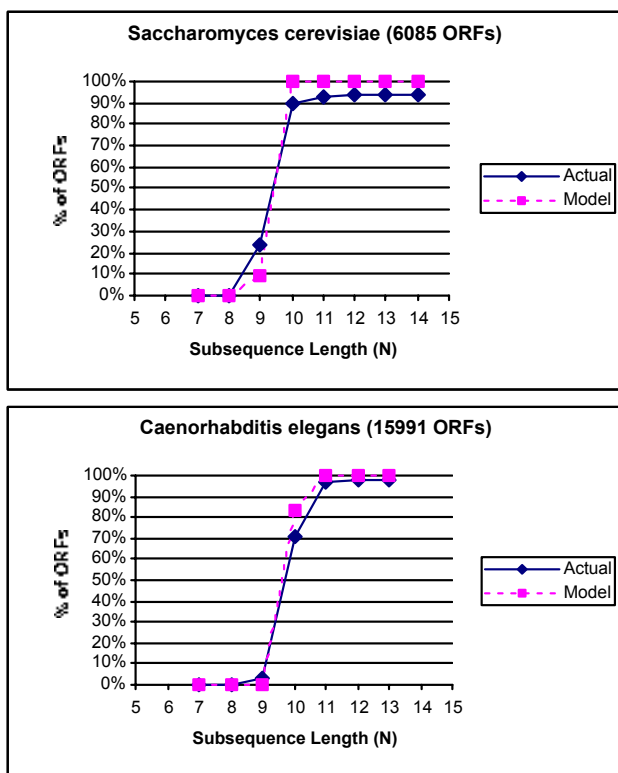


**Figure 4** Sensitivity of model (1) to ORF length, genome size of 5000 ORFs

### Results and Discussion

Fig. 5 shows plots both the predicted and actual proportion of ORFs that can be uniquely identified against probe length. Fig. 4 contains the results for four microbial species and the two eukaryotic species we studied. The appendix contains the results for all species.



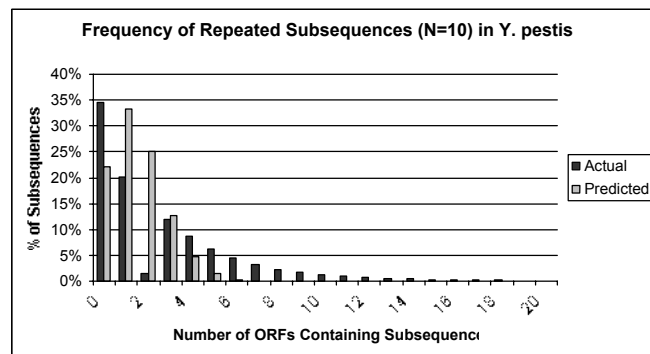


**Figure 5** Actual and predicted proportions of uniquely identifiable ORFs

The shapes of the actual and predicted plots are in close agreement, and the model predicts the exact probe length for unique identification of at least 90% of all ORFs. Thus the model, and its assumption that ORFs may be treated as random sequences, is largely correct. However, the results show that in reality, it is impossible to uniquely identify all ORFs in a genome with a short subsequence. This is further indicated by Fig. 6, a histogram showing number of subsequences of length  $N = 10$  that occur on multiple ORFs within *Y. pestis*. Contrary to the predictions of our model based on (2), some subsequences occur in as many as 20 different ORFs and more subsequences than expected do not occur in the genome at all.

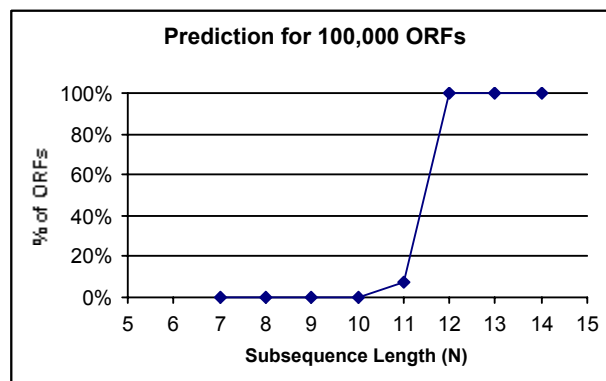
Previous work has shown that the entropy of coding regions of DNA, the ORFs, is lower than non-coding regions, indicating that they are not completely random (Schmitt and Herzel 1997; Schneider 1997). Furthermore, genomes contain detectable 10-11 base pair periodicities related to DNA folding (Herzel, Weiss, and Trifonov 1999). The characteristic sequences result in supercoiling and nucleosome formation, and common protein structure motifs. There is also evidence from whole-genome microarray experiments that inhibitory and promotional genes for the same function often have similar sequences (DeRisi, Iyer, and Brown 1997). As Fig. 4 shows, a probe length of 9 or 10 uniquely identifies almost all ORFs in a

genome. Regardless of how much longer the probes get, there will always be ORFs that can not be uniquely identified using a minimal length probe set.



**Figure 6** Truncated subsequence frequency histogram of *Y. Pestis*

Finally, Fig. 4 shows that in the case of the animal *C. elegans* at least, the simple model of random ORF sequences accurately predicts that most ORFs may be identified with a probe length of 11 bases. Fig. 7 extrapolates our model to a human-sized genome of 100,000 genes, and predicts that a probe of 12 bases is adequate to uniquely identify most human genes, although it is quite likely that a significant proportion will elude detection as demonstrated above.



**Figure 7** Random sequence model for 100,000 genes

## Conclusions

For all of the microbial species we studied, with genomes ranging from 500 to 4500 open reading frames, a probe of 9 or 10 bases is theoretically sufficient to identify at least 90% and typically 95% or more of all ORFs. Our model (1) of ORFs as random sequences of equally distributed bases and constant length is very robust: it exactly predicts the minimum required probe length from actual sequence data for all 17 prokaryotic and eukaryotic species we studied. Furthermore, the minimum probe length increases very slowly with genome size. The

model predicts that a probe of 12 bases is comprehensive enough to uniquely identify 100,000 ORFs.

Unfortunately, it is impossible to identify *all* ORFs in real sequences, contradicting the random sequence model. This suggests that a few have large nonrandom segments that they share with other ORFs. Thus, even before practical issues like hybridization error are considered, a minimum length probe set can not identify every ORF in a genome. Indeed, by using minimum length probes, we may be missing the most interesting genes with regulatory roles.

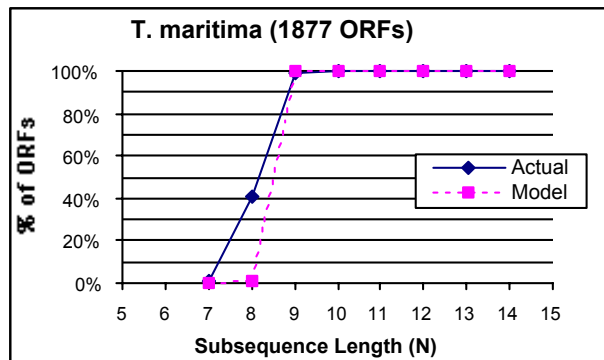
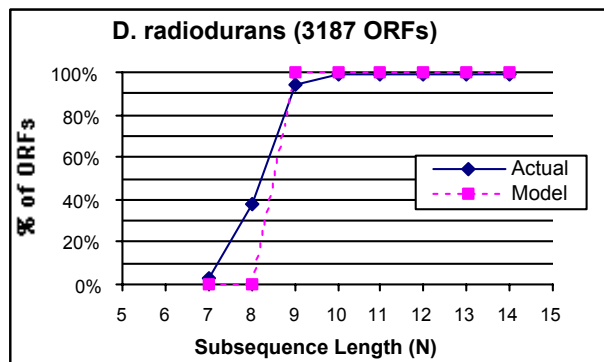
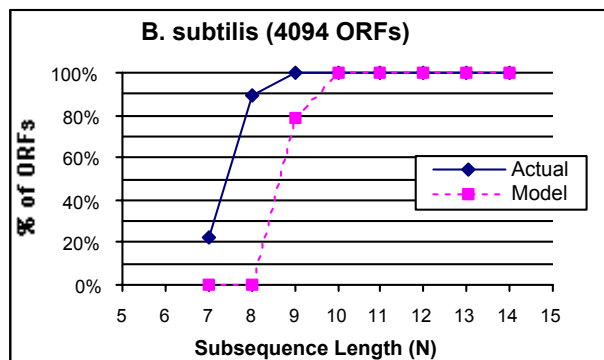
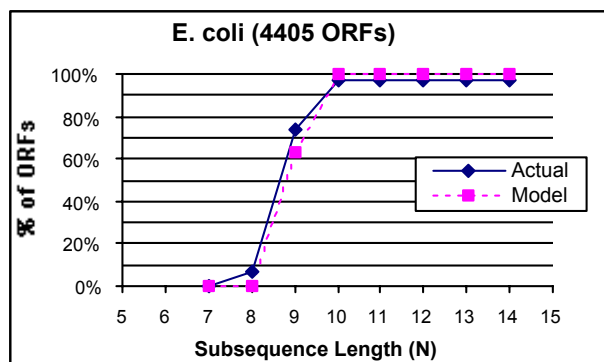
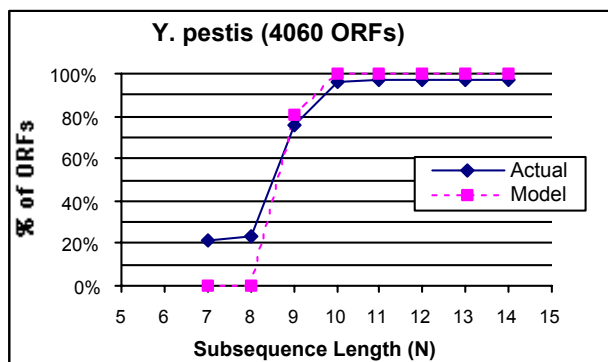
But, if measuring the expression of every gene is not critical, minimum length probes can provide significant cost and convenience advantages for “quick and dirty” whole genome expression experiments. There is no need for expensive and difficult array and probe set design, and experiments are easier to perform. It is very likely that using minimum length probes will be an important tool for inexpensive, high-throughput functional genomics.

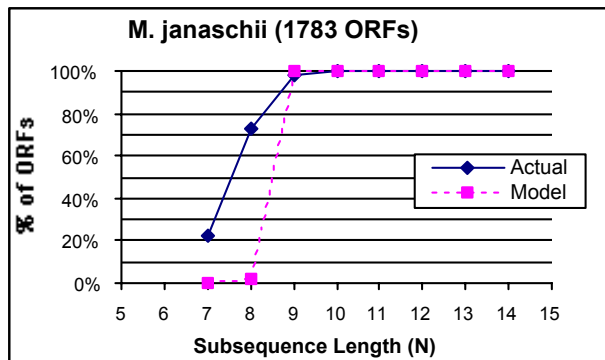
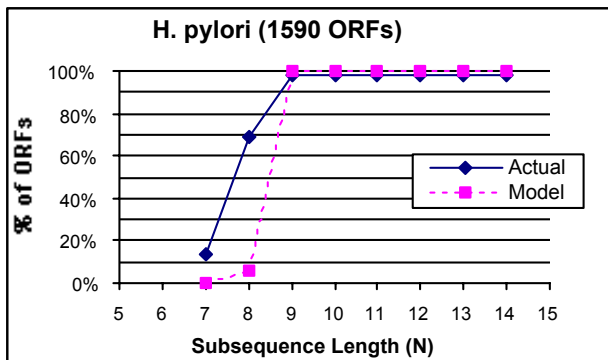
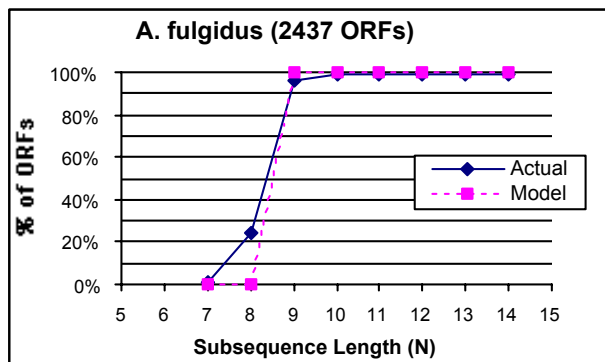
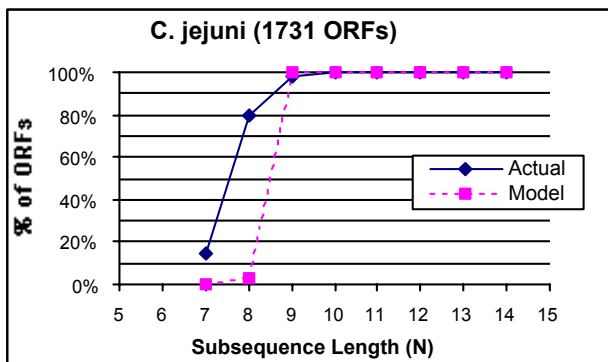
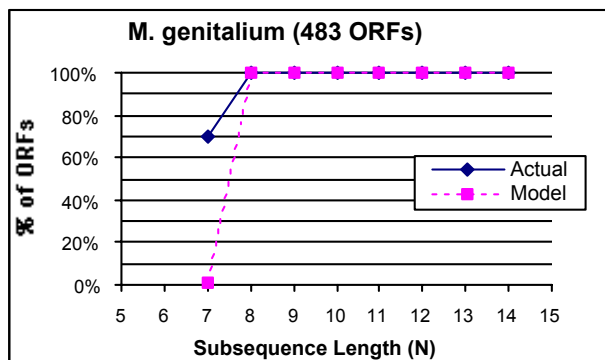
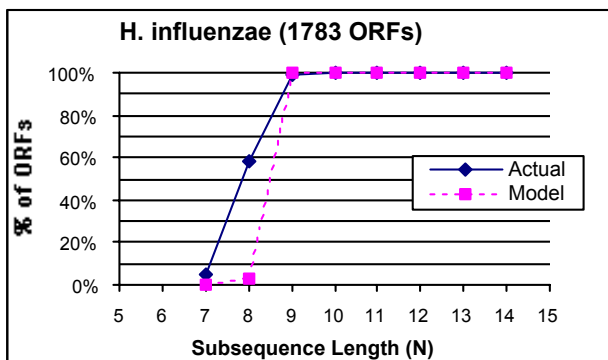
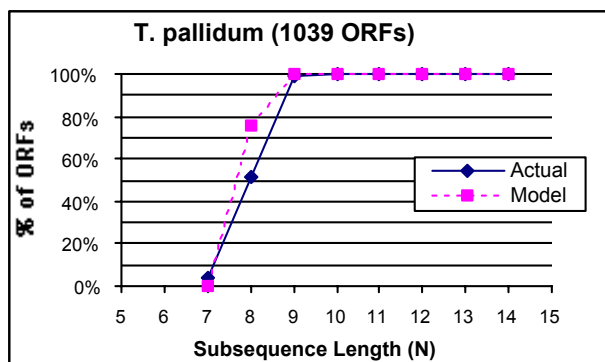
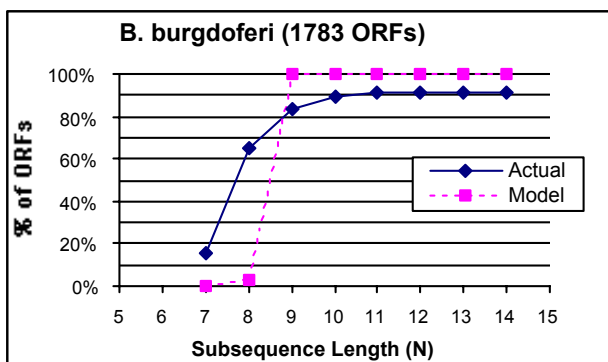
### Acknowledgements

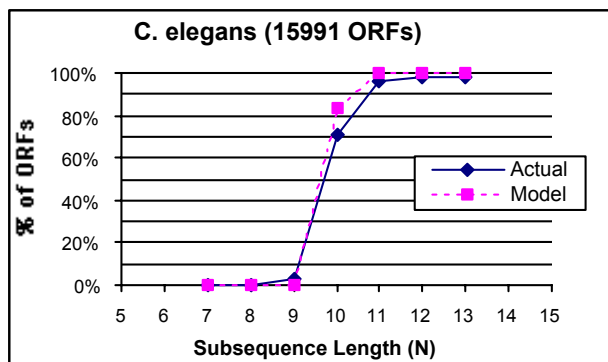
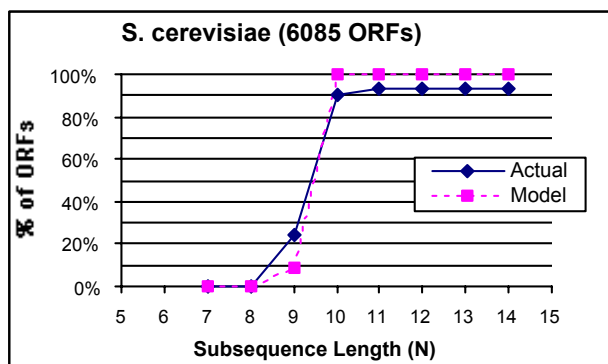
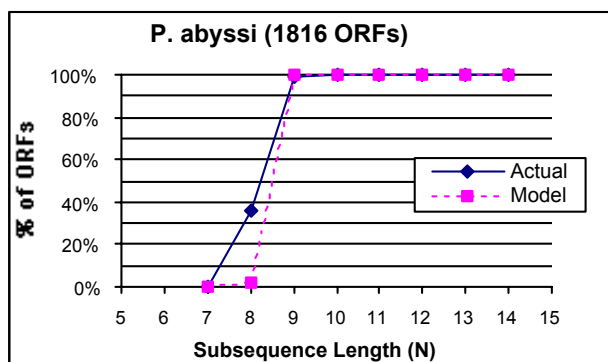
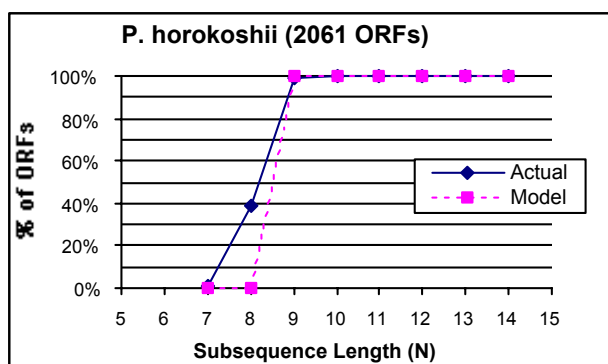
We wish to thank Tom Slezak and Paula McCreedy for several informative discussions. We also thank the gene sequencing institutions whose data we downloaded. This work was performed under the auspices of the U.S. Department of Energy by the University of California, Lawrence Livermore National Laboratory under contract No. W-7405-Eng-48.

### Appendix

As in Fig. 4, we plot the proportion of ORFs in the whole genome that are uniquely identified against probe length (N) for all species in Table 1. Shown are both analysis of real sequence data and the random sequence model prediction (1) of the number of ORFs (G), for each species in Table 1.







## References

Blattner, F. R.; Plunkett G., 3rd; Bloch, C. A.; Perna, N. T.; Burland, V.; Riley, M.; Collado-Vides, J.; Glasner, J. D.; Rode, C. K.; Mayhew, G. F.; et al. 1997. The

complete genome sequence of Escherichia coli K-12. *Science* 277:1453-1474.

Bloom, D. M. 1996. Probabilities of clumps in a binary sequence. *Mathematics Magazine* 69: 366-372.

Bult, C. J.; White, O.; Olsen, G. J.; Zhou, L.; Fleischmann, R. D.; Sutton, G. G.; Blake, J. A.; FitzGerald, L. M.; Clayton, R. A.; Gocayne, J. D.; et al. 1996. Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. *Science* 273: 1058-1073.

*C. elegans* Sequencing Consortium. 1998. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science* 282: 2012-2018.

DeRisi, J. L.; Iyer, V. R.; Brown, P. O. 1997. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* 278: 680-686.

Fleischmann, R. D.; Adams, M. D.; White, O.; Clayton, R. A.; Kirkness, E. F.; Kerlavage, A. R.; Bult, C. J.; Tomb, J. F.; Dougherty, B. A.; Merrick, J. M.; et al. 1995. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* 269: 496-512.

Fraser, C. M.; Casjens, S.; huang, W. M.; Sutton, G. G.; Clayton, R.; Lathigra, R.; White, O.; Ketchum, K. A.; Dodson, R.; Hickey, E. K.; et al. 1997. Genomic sequence of a Lyme disease spirochaete, *Borrelia burgdorferi*. *Nature* 390: 580-586.

Fraser, C. M.; Gocayne, J. D.; White, O.; Adams, M. D.; Clayton, R. A.; Fleischmann, R. D.; Bult, C. J.; Kerlavage, A. R.; Sutton, G.; Kelley, J. M.; et al. 1995. The minimal gene complement of *Mycoplasma genitalium*. *Science* 270: 397-403.

Fraser, C. M.; Norris, S. J.; Weinstock, G. M.; White, O.; Sutton, G. G.; Dodson, R.; Gwinn, M.; Hickey, E. K.; Clayton, R.; Ketchum, K. A.; et al. 1998. Complete genome sequence of *Treponema pallidum*, the syphilis spirochete. *Science* 281: 375-388.

Goffeau, A.; Barrell, B. G.; Bussey, H.; Davis, R. W.; Dujon, B.; Feldmann, H.; Galibert, F.; Hoheisel, J. D.; Jacq, C.; Johnston, M.; Louis, E. J.; Mewes, H. W.; Murakami, Y.; Philippsen, P.; Tettelin, H.; Oliver, S. G. 1996. Life with 6000 genes. *Science* 274: 546, 563-567.

Herzel, H.; Weiss, O.; Trifonov, E. N. 1999. 10-11 bp periodicities in complete genomes reflect protein structure and DNA folding. *Bioinformatics* 15: 187-193.

Kawarabayasi, Y.; Sawada, M.; Horikawa, H.; Haikawa, Y.; Hino, Y.; Yamamoto, S.; Sekine, M.; Baba, S.; Kosugi, H.; Hosoyama, A.; et al. 1998. Complete sequence and gene organization of the genome of a hyperthermophilic archaeobacterium, *Pyrococcus horikoshii* OT3. *DNA Research* 5: 55-76.

Klenk, H. P.; Clayton, R. A.; Tomb, J. F.; White, O.; Nelson, K. E.; Ketchum, K. A.; Dodson, R. J.; Gwinn, M.; Hickey, E. K.; Peterson, J. D.; et al. 1997. The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon *Archaeoglobus fulgidus*. *Nature* 390: 364-370.

Kunst, F.; Ogasawara, N.; Moszer, I.; Albertini, A. M.; Alloni, G.; Azevedo, V.; Bertero, M. G.; Bessieres, P.; Bolotin, A.; Borchert, S.; et al. 1997. The complete genome sequence of the gram-positive bacterium *Bacillus subtilis*. *Nature* 390: 249-256.

Lipshutz, R. J.; Fodor, S. P. A.; Gingeras, T. R.; and Lockhart, D. J. 1999. High Density Synthetic Oligonucleotide Arrays. *Nature Genetics Microarray Supplement*. 21: 20-24.

Nelson, K. E.; Clayton, R. A.; Gill, S. R.; Gwinn, m. L.; Dodson, R. J.; Haft, D. H.; Hickey, E. K.; Peterson, J. D.; Nelson, W. C.; Ketchum, K. A.; et al. 1999. Evidence for lateral gene transfer between Archaea and bacteria from genome sequence of *Thermotoga maritima*. *Nature* 399: 323-329.

Schmitt, A. O.; Herzel, H. 1997. Estimating the entropy of DNA sequences. *Journal of Theoretical Biology* 188: 369-377.

Schneider, T. D. 1997. Information content of individual genetic sequences. *Journal of Theoretical Biology* 188: 427-441.

Tomb, J. F.; White, O.; Kerlavage, A. R.; Clayton, R. A.; Sutton, G. G.; Fleischmann, R. D.; Ketchum, K. A.; Klenk, H. P.; Gill, S.; Dougherty, B. A.; et al. 1997. The complete genome sequence of the gastric pathogen *Helicobacter pylori*. *Nature* 388: 539-547.

Velculescu, V. E.; Zhang, L.; Vogelstein, B.; Kinzler, K. W. 1995. Serial analysis of gene expression. *Science* 270: 484-487.

White, O.; Eisen, J. A.; Heidelberg, J. F.; Hickey, E. K.; Peterson, J. D.; Dodson, R. J.; Haft, D. H.; Gwinn, M. L.; Nelson, W. C.; Richardson, D. L; et al. 1999. Genome sequence of the radioresistant bacterium *Deinococcus radiodurans* R1. *Science* 286:1571-1577.